UniCat: Crafting a Stronger Fusion Baseline for Multimodal Re-Identification - Supplementary

A Related Work

In current multimodal ReID literature, fusion approaches can be broadly categorized into late and middle fusion. Late fusion entails processing each modality independently and merging the representations in the final stages. Middle fusion, a subtype of late fusion, involves interchanging features between modalities to create an intricately fused representation upon which subsequent tasks are executed. This nuanced approach in middle fusion allows for a more integrated analysis of modalities compared to traditional late fusion.

A.1 Late Fusion Approaches:

HAMNet [4]. The paper introduces HAMNet, a multi-stream convolutional network (i.e. separate ResNet50 for each modality) tailored for multimodal vehicle ReID. It employs multiple backbones to independently process each modality; however, their method joins the features from each modality via a modified pooling operation and uses a global loss function on the joint representation, which suffers from modality laziness.

CCNet [10]. CCNet, as described in the paper, is a multimodal vehicle ReID model that employs a multi-stream architecture, with each branch dedicated to processing individual modalities. This method mirrors the approach of UniCat through its use of modality-specific loss terms. However, CCNet differs by incorporating global contrastive losses, which is suboptimal as it acts to discard modality-unique information.

PHT [6]. The paper introduces the Progressively Hybrid Transformer (PHT) for multimodal vehicle ReID. This method employs individual transformer branches for each modality and integrates them at a singular depth, while utilizing the Modal-specific Controller (MC) and Modal Information Embedding (MIE) for modality adjustments. This design embodies late fusion, given the independent modality processing and a singular fusion point using averaging. Notably, this was the previous state-of-the-art method on RGBNT100 and is the first transformer-based method to tackle the multimodal ReID problem. However, late fusion with averaging prevents each modality from extracting the maximum amount of task-relevant information per modality laziness.

A.2 Middle Fusion Approaches:

PFNET [9]. The Progressive Fusion Network (PFNet) for multimodal person ReID follows a middle fusion approach, initially processing modalities separately but later integrating them through intermediate and global stages. Despite its middle fusion characteristics, the architecture predominantly relies on simple concatenation of latent factors, which may not be optimally aligned for the desired tasks. This suggests a potential need for a more constrained inter-modality interaction and balance in learning effective representations.

IEEE [7]. The Interact, Embed and EnlargE (IEEE) method for multimodal person ReID employs independent modality processing, followed by cross-modal interactions and feature enhancements. This middle fusion strategy is evident as it interchanges and integrates modality-specific features at

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

intermediate stages. However, its reliance on basic feature summation and late-stage concatenation suggests further investigation into how to best balance the trade-offs to alleviate modality laziness.

B Training Details

For training, we combine triplet loss $\mathcal{L}_{\text{tri}}$ [3] that tries to increase class separability of the embeddings z with a cross-entropy loss $\mathcal{L}_{\text{CE}}$ that seeks to put the embeddings into class-specific subspaces [5]. Specifically, we use a modified soft-margin triplet loss that includes hard-mining [8] to select the most challenging positive and negative samples (p and n) for a given anchor (a):

$$\mathcal{L}_{\text{tri}} = \min_{n} \max_{p} \log\left(1 + \exp\left(||a - p||_2 - ||a - n||_2 + \alpha\right)\right) \tag{1}$$

To implement the hard-mining, we select for each anchor in a given mini-batch the positive/negative sample in the same mini-batch that is most farthest/closest to the anchor in the latent space. For the cross-entropy (CE) loss, class-wise output scores are found via a linear classifier, and then compared against the ground truth. When combining these losses, we set the balancing coefficient λ to 1.

Images are resized to 128x256 for RGBNT100/RGBN300 and 256x128 for RGBNT201. During training, images are augmented with random horizontal flipping, padding, random cropping and random erasing [3]. We use many of the training strategies suggested by [5], including a linear warmup strategy for the learning rate and using BNNeck to normalize features before CE loss. For each fusion strategy, we perform grid-search hyperparameter optimization across three batch sizes ([64, 128, 256]) and three learning rates ([.008, .016, .032]). We train for 120 and 200 epochs for ViT-B and ResNet-50 backbones, respectively, used SGD with momentum 0.9 and a cosine learning rate decay. All backbones are instantiated with Imagenet [1] pre-trained weights while the linear classifier and bottleneck layers are instantiated via [2]. Similar to [5], we choose cosine distance as our similarity metric for ReID.